Sonority Measurement using System, Source and Suprasegmental Information

Bidisha Sharma and S. R. Mahadeva Prasanna, *Senior Member, IEEE*

***Abstract*—Sonorant sounds are characterized by regions with prominent formant structure, high energy and high degree of periodicity. In this work, the vocal-tract system, excitation source and suprasegmental features derived from the speech signal are analyzed to measure the sonority information present in each of them. Vocal-tract system information is extracted from the Hilbert envelope of numerator of group delay function. It is derived from zero time windowed speech signal that provides better resolution of the formants. A five-dimensional feature set is computed from the estimated formants to measure the prominence of the spectral peaks. A feature representing strength of excitation is derived from the Hilbert envelope of linear prediction residual, which represents the source information. Correlation of speech over ten consecutive pitch periods is used as the suprasegmental feature representing periodicity information. The combination of evidences from the three different aspects of speech provides better discrimination among different sonorant classes, compared to the baseline MFCC features. The usefulness of the proposed sonority feature is demonstrated in the tasks of phoneme recognition and sonorant classification.**

***Index Terms*—Sonority, phoneme recognition, source, system, suprasegmental, zero time windowing.**

I. Introduction

Sonority refers to relative loudness of speech sounds [1]. Most of the sonorant sounds are produced using relatively less constricted vocal-tract shape and glottal vibration. This results in regions of regular structure having high energy and high degree of periodicity. The sonorant regions are therefore prominent ones in the speech signal and important for many speech processing tasks [2]. Vowels are the most sonorous sounds, which mostly form the nucleus of a syllable. Different sonority hierarchies are defined in the literature as mentioned in [1]. However, the most commonly referred sonority hierarchy for the six major classes of sonorants in the decreasing order of sonority is *low-vowels, mid-vowels, high-vowels, glides, liquids and nasals*. In [3], the sonority hierarchy for obstruents is defined in the decreasing order of sonority as *voiced fricatives, voiced affricates, voiced stops, voiceless fricatives, voiceless affricates, and voiceless stops*.

Sonority is used to explain both the perception of syllables and their phonetic structure [4]. The *sonority sequencing principle* states that in every syllable, syllable nucleus has the highest sonority value [5]. According to *syllable contact law*, the junction between two syllables is well recognizable when the coda of the present syllable has higher sonority value than the onset of the next syllable [6]. According to [7], the syllables with nuclei having more sonority value tend to have more stress compared to the syllables with nuclei having less sonority value. For example, syllables with [e] or [o] may be perceived as having more stress than those with [i] and [u]. The possible sequence of consonants present in the syllable onset and coda also depends on the sonority value associated with them. For example, consonant clusters present in syllable onset of the form [pl], [dr], [km] are very common, but the reverse order is rare. In this case, [l], [r], [m] are more sonorous than [p], [d], [k]. Similarly, [mp] and [nd] are very common as syllable codas than [pm], [dn], where [m], [n] are more sonorous than [p], [d]. Therefore, sonority of a sound unit has an impact on the basic production pattern of speech sounds. In several studies of phonology such as consonant cluster, sonorant-obstruent cluster, syllable onset and coda position, degree of sonority is used [8], [9]. *Degree of sonority* can be defined as sequential variation in various attributes that correlate to sonority, with respect to distinctive category of sound units. The variation in degree of sonority associated with different sound units is due to the change in the behavior of different articulators during production. This is also manifested in the produced speech signal.

A. Production aspects of different sonorant sounds

The most sonorant sounds, vowels, are produced with less constricted vocal-tract configuration through manipulation of the vocal-tract between glottis and lips. Position and configuration of different articulators has effect on the spectrum of generated speech signal. Narrowing the cross sectional area in the front part of vocal-tract and widening towards the back results in the decrease of first formant frequency (F_1). As a consequence of variation in position and length of constriction, second formant frequency (F_2) changes for different category of sonorants. The bandwidth of formant is associated with loss in the vocal-tract. Thus with the increase in sonority, the vocal-tract constriction decreases that results in increase in F_1, F_2 and decrease in formant bandwidth.

Compared to the obstruents, sonorants have sufficient opening of the vocal-tract to produce voicing and well defined prominent formant structure [10]. Looking into these aspects of sonorant sounds, it is expected that, accurately estimating the vocal-tract spectrum (VTS) and analyzing the formant structure may be helpful to characterize the change in vocal-tract shape with the change in degree of sonority. Due to the glottal open and closed phase, the formant structure does not show a constant behavior during one pitch period [11], [12]. The characteristics of the vocal-tract system in the open phase varies due to the coupling with vocal-fold and trachea. Whereas, during the closed phase, the speech signal

Bidisha Sharma and S. R. Mahadeva Prasanna are with the Dept. of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, India. This work is part of ongoing project on development of text to speech synthesis systems in Indian languages. Email: {s.bidisha, prasanna}@iitg.ernet.in

is mainly due to free resonances since there is no coupling with trachea and vocal-folds [13]. Therefore, extraction of VTS from speech signal corresponding to the closed phase of each pitch period may give accurate formant estimation along with it's associated measures. But, in voiced region, the glottal closing is abrupt and the duration of the closed phase is smaller than that of the open phase. For extracting the VTS, processes like linear prediction (LP) analysis and short time Fourier transform (STFT) involve block processing and are dependent on the size and position of window. Also, these methods mask the changing shape of the vocal-tract and give an average spectrum [13].

In this work, Hilbert envelope of numerator of group delay function (HNGD) spectrum derived from speech signal around the glottal closure instant (GCI) is used to estimate the VTS [14]. The GCI locations are estimated using the zero frequency filtered (ZFF) signal [15], as it is found to be more robust compared to other state-of-the-art techniques [16]. A highly tapering window is used to emphasize the speech samples around each GCI that correspond to the glottal closed phase. The sonority information present in the VTS is extracted using knowledge from the first three formants of the HNGD spectrum.

With change in the vocal-tract constriction, there is also an effect on the amplitude and spectrum of the source. Due to the change in constriction, there is fluctuation in supra-glottal pressure which has an impact on the pressure inside the glottis during the open phase of glottal vibration. This changes mechanical motion of the vocal-folds. The net effect is reduction in the amplitude of glottal source which is reflected in the Hilbert envelope (HE) of LP residual as strong peaks. These peaks have correlation with an acoustic feature called strength of excitation (SoE) as discussed in [17]. With the increase in degree of sonority, SoE also increases. Hence, it can be hypothesized that, deriving an adequate representation of SoE may add some advantage in deriving sonority information from the speech signal.

Along with the change in behavior of the vocal-tract system and the excitation source with degree of sonority, temporal variation in the speech signal also takes place. This can be observed over several pitch periods. One such measure is periodicity, which is tendency of the signal to repeat similar structure over several pitch periods. This occurs, since human speech production system changes in a continuous manner. During the production of sonorant sounds, the vocal-tract shape changes slowly and hence maintains periodicity over longer duration compared to other sounds [18]. This suprasegmental behavior of sonorants is not taken into account while analyzing vocal-tract system and excitation source perspectives. Hence, examining the regularity in the signal structure or correlation over several small segments of the speech signal may be helpful to obtain feature representing this aspect of sonority.

B. Usefulness of sonority feature

Deriving sonority feature from speech signal may be helpful in many speech processing tasks. These include, but are not limited to detection of syllable nucleus, vowel onset point detection, phoneme classification, study of syllable structure and syllabification in different languages. Sounds with higher degree of sonority form syllable nucleus. It gives information about number of syllables present in the speech signal. Number of syllables divided by duration of the signal defines syllable rate/speaking rate. There are several approaches in the literature towards this direction. In [19], syllable nucleus is detected by loudness estimation. Energy peaks in the frequency range from 250 - 2500 Hz have good correlation with syllable nuclei. Many other methods use vowel recognizer to find syllable nucleus as given in [20].

Correlation between prominent subbands is used to capture well defined formant structure in the syllable nuclei in [21]. Before applying cross-correlation between subband energy vectors, frames are weighted by Gaussian window and then temporal correlation is estimated in order to retain inter-syllable discontinuity in case of fast speech. Then, thresholding and pitch validation of subband correlation envelope is performed to enhance the detection of syllable nucleus. In the same work, experiments are also performed to find syllable nuclei which include sonorant sounds other than vowels. The mean error calculated is more in this case. This proves that the feature cannot detect all sonorant sounds. In [22], perceptually significant evidences such as excitation source peaks in LP residual and formant peaks which contribute to the loudness are used to find the most sonorous region within syllable. All these efforts are aimed to detect basically the most sonorous sounds, the vowels. There are many confusions reported within the sonorants (vowels, glides, liquids, nasals) while detecting the vowels.

Segmentation of speech into sonorant regions with high accuracy is essential for applications like automatic speech recognition (ASR) to detect the regions with high signal to noise ratio (SNR) in the speech signal [23]. In literature, sonorant segmentation is performed by using mel frequency cepstral coefficients (MFCCs), knowledge based acoustic features or a combination of both [2], [24]. Recently in [23], [25], features based on both spectral and source information are proposed and a hierarchical algorithm is developed to detect sonorant and non-sonorant regions in continuous speech. However, the feature may not have potential to further divide the sonorant regions based on the degree of sonority associated with the sound. In order to improve the performance of sonority detection, it is important to first quantify the degree of sonority associated with different sound units in a given speech segment, without having knowledge of phone sequence. In this work, an evidence is obtained which represents instantaneous sonority i.e. continuous change in sonority with time in the speech signal. In traditional methods, sonority is derived from the phone identity knowledge.

Looking into these studies present in the literature, it can be considered important to derive some feature which represents degree of sonority from speech signal. In this work, three different aspects of speech signal, namely vocal-tract system, excitation source and suprasegmental are analyzed to extract prospective features to discriminate among different classes of sonorants. The three attributes are analyzed individually and effectively combined to derive a multi-dimensional feature

which can represent sonority. The obtained sonority feature is used in phoneme recognition and results show improvement. In the analysis of all features, focus is on classifying within the sonorants according to the sonority hierarchy.

Rest of the paper is organized as follows: Features of vocal-tract system for sonority detection are proposed in Section II. Features of excitation source and suprasegmental feature are presented in Section III and Section IV, respectively. Section V describes the combination of proposed evidences to represent sonority measure. Section VI shows the experiments performed to demonstrate the usefulness of sonority evidence in different speech processing task such as phoneme classifier. In Section VII, summary, conclusions and future direction are mentioned.

II. Features of vocal-tract system for Sonority Detection

The categorical formant structure in the VTS of sonorant sounds can be interpreted by measures associated with amplitude of spectral peaks and valleys, formant bandwidths and slope. Bandwidth of the spectral peak decreases, while the spectral peak value increases with increase in degree of sonority. The peak-to-valley ratio (PVR) of spectral peak is a direct representation of spectral prominence, that is inversely proportional to the corresponding bandwidth. Spectral prominence refers to spectral peaks with more sharpness and higher energy, which increases with degree of sonority. This depends on PVR, slope, bandwidth and amplitude associated with spectral peaks. Narrow constriction results in relatively low values of formant frequencies and spectral peaks. High-vowels are produced by raising the tongue body thus forming narrow constriction in the front part of vocal-tract. This results in decrease in F_1 and increase in bandwidth, primarily due to acoustic losses in the vocal-tract walls and glottis. As explained in [26], due to less spacing between F_1 and F_0, the response of low frequency auditory nerve fibers are dominated in low frequency region by F_1, resulting in production of relatively stable response in auditory system. In contrast to high-vowels, low-vowels are produced by narrowing the posterior part and widening towards lips, resulting in increase in F_1 and higher difference between F_1 and F_0. Due to this difference, the auditory nerve fibers near F_0 are not dominated by F_1. As a consequence, there is a fall in the spectrum below F_1 [26]. Due to the intermediate position of tongue body during production of mid-vowels, F_1 also lies in between that of high-vowel and low-vowel. In this case, the auditory nerve fibers are in synchrony with either F_1 or F_0. Fluctuation of second and third formant frequencies, F_2 and F_3 depends on the constriction length and position in the vocal-tract.

During the production of nasals, the vocal-tract is completely closed, while the velopharyngeal part is open and there is no pressure increase behind the constriction. In this case, during the time of closure of vocal-tract, if the vocal-folds are in a position of voicing, the same will continue after the closure [26], [27]. Nasals have the first formant at a very low frequency and with less energy. The higher formants are also of weak amplitudes. Glides are produced by forming narrow constriction to an extent, so that there is no significant pressure drop across the constriction. This results in vibration of vocal-folds and lower F_1 with wider bandwidth. As an influence of the narrow constriction, the glottal source also gets modified. The liquids are also produced with narrow vocal-tract constriction, but the length of the constriction is shorter than that of the glides. As a consequence, F_1 of liquids is higher than that of glides. During production of liquids, the tongue is shaped in such a way that there is a split in the vocal-tract, which cannot be compared with an uniform tube [26].

With the increase in vocal-tract constriction, F_1 decreases and bandwidth of first formant increases gradually along the sequence of following sounds: low-vowels, mid-vowels, high-vowels, liquids, glides and nasals. With decrease in F_1, there is significant reduction in the overall spectrum amplitude. Amplitude of F_2 is dependent on F_1 and on the point of constriction along the vocal-tract. Since sonority associated with a sound unit depends on the vocal-tract constriction, the process for extraction of VTS should be appropriate.

A. HNGD Spectrum

HNGD is found to have potential in deriving VTS for a very short segment of speech signal around GCI that mostly corresponds to the glottal closed phase as reported in [14]. It is employed in this work to analyze different characteristics of VTS for sonorant sounds. The same process of deriving HNGD spectrum around each GCI in the speech signal, as in [14] is used here:

- The frequency response of ZFF as proposed in [15] can be represented by (1). The analogous time domain window function shown in (2) is used to emphasize the speech samples closest to each GCI location. This windowing method is referred as zero time windowing (ZTW) [14].

$$|H(w)| = |1/(1-z^{-1})^2|_{z=e^{jw}} = 1/2(1-cosw) \\ = 1/4sin^2(w/2) \quad (1)$$

$$w[n] = \begin{cases} 0 & n = 0; \\ 1/(4sin^2(\pi n/(2N))) & n = 1, 2, ...N-1. \end{cases} \quad (2)$$

  where, N is the length of the window.
- Let $s(n)$ be the speech signal and corresponding epoch locations are extracted by using ZFF signal as explained in [15]. This can be represented by a train of impulses as shown in (3), where M is total number of epochs and i_k is the estimated epoch location [28].

$$\sum_{k=1}^{M} \delta(n - i_k) \quad (3)$$

- Let $x_k(n)$ be the windowed signal derived by placing the window at each epoch location as shown in (4)

$$x_k(p) = s(p) \times w(n) \quad (4)$$

  where, $p = i_k, i_k + 1, ...i_k + N - 1$ and N is length of window function ($w(n)$).

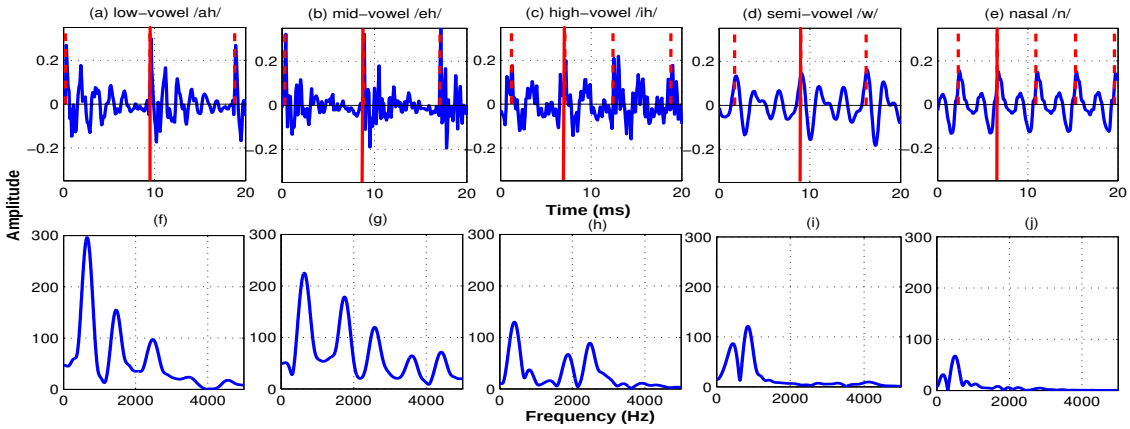

Fig. 1. *HNGD spectra for different classes of sounds showing apparent discrepancy in the spectrum shape. First row depicts* 20 *ms segment of (a) low-vowel /ah/, (b) mid-vowel /eh/, (c) high-vowel /ih/, (d) semi-vowel /w/, (e) nasal /n/ from TIMIT test database with dashed vertical lines representing epoch locations. Second row (f), (g), (h), (i), (j) show corresponding HNGD spectra, respectively, for* 5 *ms segment around the epoch location represented by solid line.*

- Due to highly decaying nature of the window function, there is possibility of masking of formant peaks by over-smoothing and thereby loosing required evidences from formants. This effect of peaks merging or smoothing can be avoided by using Fourier transform phase spectra i.e. group-delay (GD) spectra instead of usual magnitude spectra [29]. The numerator of the GD function (NGD) ($g(w)$) of $x_k(n)$ is computed as in [14]

$$g(w) = X_R(w)Y_R(w) + X_I(w)Y_I(w) \qquad (5)$$

  where, $X(w) = X_R(w) + jX_I(w)$ is the discrete time Fourier transform (DTFT) of $x_k(n)$ and $Y(w) = Y_R(w) + jY_I(w)$ is the DTFT of $y_k(n) = nx_k(n)$. The subscripts 'R' and 'I' denote real and imaginary parts, respectively.
- The spectral resolution is enhanced by successively differentiating NGD two times (DNGD), which shows sharp peaks at each formant location.
- In order to highlight these peaks further, HE of the DNGD is computed which is called HNGD spectrum.

For different categories of sound units, HNGD is found to have the potential to detect formant characteristics with accuracy for short window, as reported in [14]. This motivate to exploit usefulness of HNGD spectrum in characterizing VTS to derive sonority feature.

B. Effectiveness of HNGD spectrum for sonority detection

In order to substantiate the variation in formant structure of the HNGD spectra with respect to degree of sonority, the same is shown in Fig. 1 for different classes of sounds. Figures 1 (a) - (e) show 20 ms segments of low-vowel /ah/, mid-vowel /eh/, high-vowel /ih/, semi-vowel /w/, nasal /n/, respectively. The epoch locations marked with dashed vertical lines are derived using ZFF method as described in [15]. Figures 1 (f) - (j) show HNGD spectra around the epochs represented by solid lines in Fig. 1 (a) - (e), respectively. For the spectrum of low-vowel /ah/, first three spectral peaks have higher amplitudes, higher slopes and lower bandwidths. The slope represents rate of decay of the spectrum amplitude from

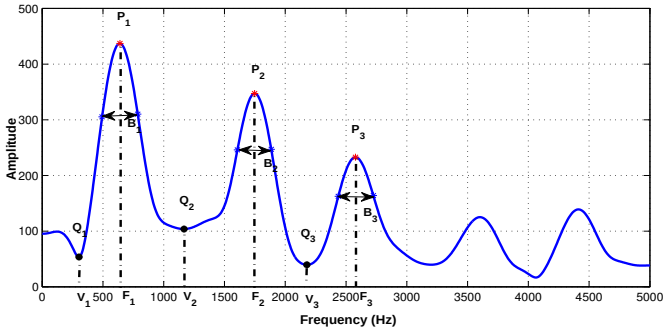

Fig. 2. *Vocal-tract spectrum represented by HNGD spectrum corresponding to /eh/ showing different measurements i.e. first three formant frequency values (in Hz), amplitude of spectral peaks, frequency at spectral valleys (in Hz), amplitude of spectral valleys and bandwidth*

spectral peaks (P_1, P_2, P_3) to corresponding preceding spectral valleys (Q_1, Q_2, Q_3) as in Fig. 2. On the other hand, mid-vowel /eh/ has lower F_1 and hence lower spectral prominence than that of the low-vowel. For high-vowel, F_1 decreases further. With the decrease in F_1 value, reduction in overall spectrum amplitude can also be observed. For semi-vowel and nasal sounds, differences between different attributes of spectra are depicted in Fig 1(i) and (j). Influenced by these observations, sonority feature is represented using different statistics from the HNGD spectrum.

C. Proposed features of vocal-tract system to find degree of sonority

In order to find the degree of sonority associated with a sound unit, different attributes of VTS are derived from the HNGD spectrum, obtained around each epoch location. Different classes of sonorant sounds from TIMIT database used in this study are *nasals* ([m], [n], [ng]), *liquids* ([r], [l]), *glides* ([w], [y]), *high-vowels* ([ih], [iy], [uh], [uy]), *mid-vowels* ([eh], [ey], [oy], [ow]) and *low-vowels* ([aa], [ah], [ae]). These categories of sound units are segmented according to the information in TIMIT label files, succeeded by normalization with respect to the maximum value of each sound unit,

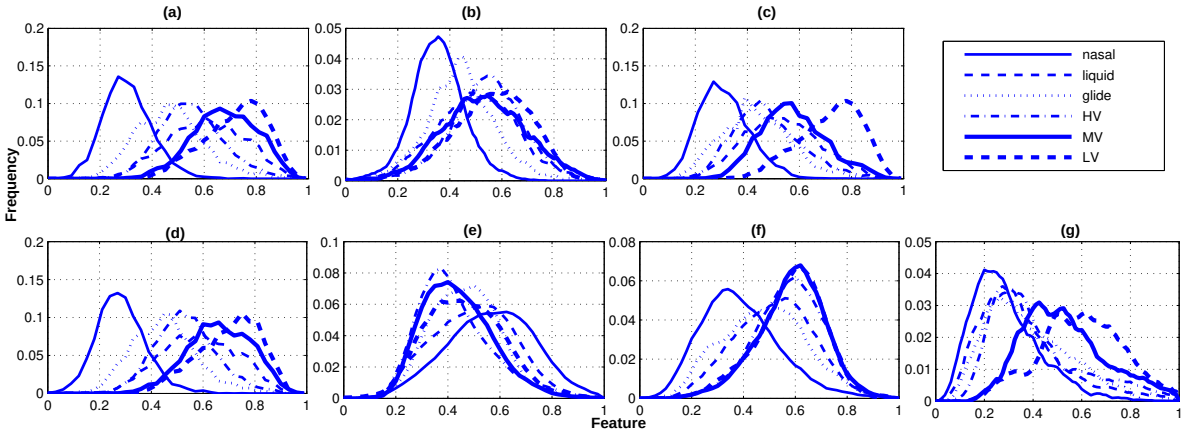

Fig. 3. *Distributions of the proposed sonority features for different sonorant sound units. Distribution for feature (a)f_1, (b)f_2, (c)f_3, (d)f_4, (e)f_5, (f) feature of excitation source (f_6) and (g) suprasegmental feature (f_7)*

for which epoch locations are derived. HNGD spectrum of energy normalized speech segment after each epoch location, is obtained as described in Section II-A which has potential to correctly characterize VTS [14]. The first three formant frequencies and associated measures are of crucial importance in many speech processing studies. Therefore, the same in HNGD spectra are employed for the task of extraction of features having capability to represent sonority. The effectiveness of each of the proposed features can be justified from the distribution curves obtained for the entire TIMIT test database for different classes as shown in Figure 3.

Following measures are extracted from the estimated VTS for measuring sonority.

1) Formant peak values: The first three formant frequency values (in Hz) obtained from HNGD spectrum are F_1, F_2, F_3 and the corresponding amplitude of spectral peaks are represented by P_1, P_2, P_3 as shown in Fig. 2. With the increase in degree of sonority, F_1 (in Hz) increases. This is also reflected in the amplitude of spectral peaks, as increase in F_1 results in overall increase in the spectrum amplitude. The mean amplitude of first three spectral peaks is calculated, which is represented as f_1, where, $f_1 = \frac{1}{3}\sum_{i=1}^{3} P_i$. The estimated distribution of normalized value of f_1 for different classes of sonorant sounds is shown in Fig. 3(a). It can be observed from Fig. 3(a) that f_1 may not discriminate well between different sonorant classes, but it does provide some evidence along the lines of sonority hierarchy.

2) Formant peak deviation: When two or more formant frequencies come close together, there is an increase in spectrum value in the vicinity of these formant frequencies. The next measure for sonority measurement from VTS is the mean of relative deviation between amplitude of first three spectral peaks. Here D_1 and D_2 are differences between amplitudes of first and second spectral peaks, and second and third spectral peaks, respectively. The mean of these differences is represented as $f_2 = \frac{1}{2}\sum_{i=1}^{2} D_i$. The distribution corresponding to normalized value of f_2 for different sonorant classes derived from whole TIMIT test database is shown in Fig. 3(b). f_2 may provide some information along the sonority hierarchy.

3) Spectral valleys preceding the first three formant peaks: Along with spectral peaks, spectral valleys are also of importance for overall study of the spectrum shape. Spectral valleys (V_1, V_2, V_3) preceding to the first three formant frequencies (F_1, F_2, F_3) are detected and the mean value of corresponding spectral amplitudes Q_1, Q_2, Q_3 is calculated. It is represented as $f_3 = \frac{1}{3}\sum_{i=1}^{3} Q_i$. The distribution of normalized f_3 derived from segments of different sonorant classes from entire TIMIT test database is shown in Fig. 3(c).

4) Slope associated with each formant peak: In order to detect spectral prominence, slope associated with each spectral peak is also measured. To measure the slope, first three spectral peaks (P_1, P_2, P_3) corresponding to formant frequency values F_1, F_2, F_3 are detected. Similarly, preceding amplitude of spectral valleys (Q_1, Q_2, Q_3) corresponding to frequency values V_1, V_2, V_3 are determined as shown in Fig. 2. Then, slope associated with each of the first three spectral peaks is calculated as follows:

$$SP_1 = \frac{P_1 - Q_1}{F_1 - V_1}; SP_2 = \frac{P_2 - Q_2}{F_2 - V_2}; SP_3 = \frac{P_3 - Q_3}{F_3 - V_3} \quad (6)$$

To represent this feature, average value of SP_1, SP_2 and SP_3 is calculated as, $f_4 = \frac{1}{3}\sum_{i=1}^{3} SP_i$. The distributions are obtained for normalized value of f_4 for different sonorant classes in the TIMIT test database as shown in Fig. 3(d).

5) Formant Bandwidth: Formant bandwidth is directly proportional to the loss associated with vocal-tract. This may arise from different sources such as vocal-tract walls, viscosity, heat conduction and radiation. Hence, with more constricted vocal-tract configuration, bandwidth associated with peaks also increases. This results in decrease in degree of sonority. Before calculating the bandwidth, the spectrum is converted to log scale $(10\log(g(w)_{hngd}))$, where, $g(w)_{hngd}$ represents HNGD spectrum. For each of the first three spectral peaks (P_1, P_2, P_3), corresponding 3 dB bandwidths (B_1, B_2, B_3) are measured and average bandwidth is calculated ($f_5 = \frac{1}{3}\sum_{i=1}^{3} B_i$). The distributions corresponding to normalized bandwidth is shown in Fig. 3(e), which decreases with the increase in sonority.

The values of each of the features f_1, f_2, f_3, f_4, f_5 obtained from all the frames across all instances of the six types of

sounds are normalized as follows:

$$f_i = \frac{f_i - min(f_i)}{max(f_i) - min(f_i)} \tag{7}$$

where, i ranges from 1 to 5. $min(f_i)$ and $max(f_i)$ represent minimum and maximum values of f_i extracted over all classes of sonorant sounds for entire TIMIT test database.

D. Combined Vocal Tract feature to find degree of sonority

TABLE I
Canonical correlation analysis (CCA) between different features of vocal-tract system

Features	Correlation value
$\mathbf{f_1}$ and $\mathbf{f_2}$	0.89
$\mathbf{f_1}$ and $\mathbf{f_3}$	0.88
$\mathbf{f_1}$ and $\mathbf{f_4}$	0.63
$\mathbf{f_1}$ and $\mathbf{f_5}$	0.40
$\mathbf{f_2}$ and $\mathbf{f_3}$	0.89
$\mathbf{f_2}$ and $\mathbf{f_4}$	0.52
$\mathbf{f_2}$ and $\mathbf{f_5}$	0.38
$\mathbf{f_3}$ and $\mathbf{f_4}$	0.59
$\mathbf{f_3}$ and $\mathbf{f_5}$	0.39
$\mathbf{f_4}$ and $\mathbf{f_5}$	0.33

Each of the features f_1, f_2, f_3, f_4 and f_5 are normalized and approximated by Gaussian probability density function as shown in Fig. 3 (a), (b), (c), (d), (e), respectively. The distributions do not provide clear discrimination among different classes of sonorants. However, still the increasing trend of the features f_1, f_2, f_3 and f_4 from nasals to low-vowels can be observed, while f_5 exhibits a decreasing trend for the same. Also, some disparity in terms of overlap of distributions among different classes of sounds for each of the features of VTS can be interpreted from Fig. 3 (a)-(e). For example, in the distribution of f_2, a distinct overlap between the low-vowel, mid-vowel and high-vowel can be observed. f_1 shows less overlap between the three vowel categories along the line of sonority hierarchy. f_2 has lower amount of overlap between the distributions of glides and nasals.

It can be inferred from Fig. 3(c) that, f_3 possess better adequacy to bring out the differences between low-vowel and mid-vowel compared to other features. In each of f_1, f_3 and f_4, the liquids have higher values than that of glides, whereas according to the sonority hierarchy, glides are more sonorous than the liquids. In Fig. 3(e), f_5 shows a correct reverse trend of feature values with respect to the sonority hierarchy. However, the extent of overlap between different classes is more compared to other features. Based on this interpretation, it can be inferred that the five derived features of vocal-tract system may carry different information.

The redundancy among the five attributes derived from the VTS is elucidated using canonical correlation analysis (CCA) [30], [31]. The correlation values derived from CCA among different pairs of features are shown in Table I. Although correlation exists between the five features of vocal-tract system, there is some extra information captured by each feature, as the correlation value is less than 1 in each case. Based on these observations, a five-dimensional feature vector of vocal-tract system is proposed in this work, which has the ability to quantify the sonority hierarchy.

III. Excitation source information for sonority detection

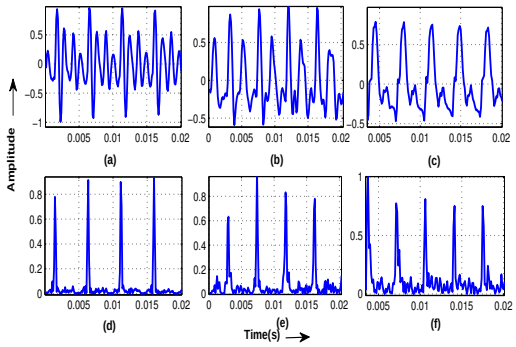

Fig. 4. *Illustration of difference in nature of excitation source in vowels, semi-vowels and nasals. (a)-(c) show* 20 *ms speech segment of vowels, semi-vowels and nasals. (d)-(f) show corresponding HE of LP residual, respectively.*

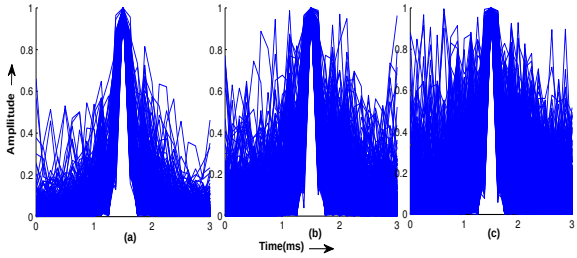

Fig. 5. 3 *ms duration of superimposed segments of HE of LP residual in the vicinity of impulse-like excitations for (a) Vowels, (b) semi-vowels, (c) nasals.*

Strength of excitation (SoE) is related to the abruptness of glottal closure, which is maximum for an ideal impulse and corresponds to strength of differenced electro-glotto-graph (DEGG) signal at GCIs. In order to visualize how SoE changes with degree of sonority, an effective representation of SoE derived from excitation source needs to be explored. Given the speech segment of particular sound unit (vowels, semi-vowels or nasals), LP analysis can be performed to derive the LP coefficients. The residual signal is obtained by inverse filtering the speech signal using LP coefficients. The inverse filtering suppresses the vocal-tract characteristics from the speech signal and mostly contains information about the excitation source. The residual signal shows noise like characteristics in unvoiced regions and large discontinuity in voiced regions of the speech signal. This is a good approximation of excitation source signal when LP order is properly chosen [32]. In this work, the LP residual is derived by performing LP analysis on overlapped segments of speech signal (size of frame =25 ms, frame shift = 5 ms, LP order = 10 and sampling frequency = 8 kHz). The GCIs are manifested as large amplitude fluctuations, either in positive or negative polarity in the LP residual.

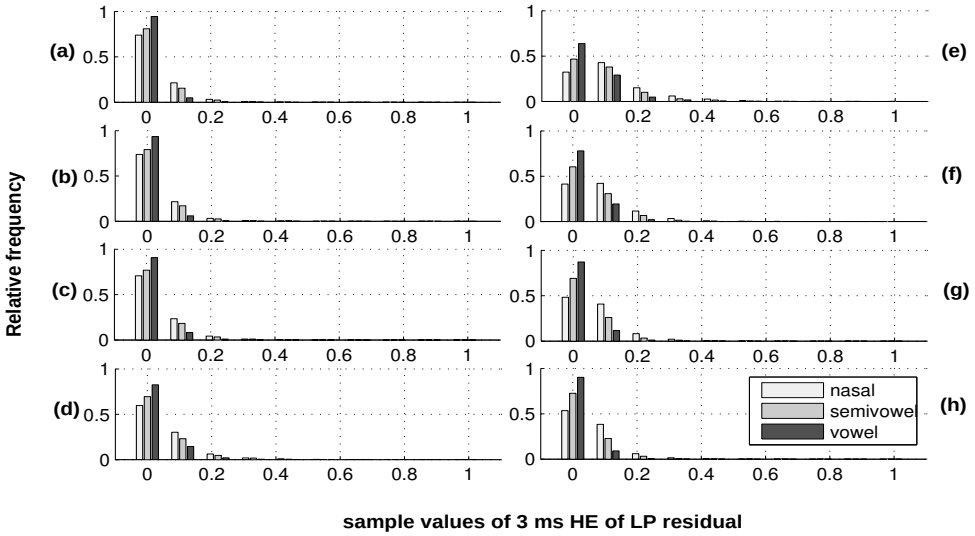

Fig. 6. *Histogram plot of sample values of* 3 *ms HE of LP residual.* 3 *ms segment is divided into* 0.25 *ms frames. (a), (b),(c),(d) correspond to* 0 *to* 1 *ms and (e), (f), (g), (h) corresponds to* 2 *to* 3 *ms of the* 3 *ms segment*

This difficulty can be overcome by using the HE of LP residual [33]. The HE $h_e(n)$ of LP residual $e(n)$ is defined as

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{8}$$

where, $e_h(n)$ is Hilbert transform of $e(n)$ and in given by

$$e_h(n) = IDFT[E_h(k)] \tag{9}$$

where,

$$E_h[k] = \begin{cases} -jE(k) & k = 0, 1, ...(\frac{N}{2}) - 1; \\ jE(k) & k = (\frac{N}{2}), (\frac{N}{2}) + 1, ...(N-1) \end{cases} \tag{10}$$

IDFT denotes inverse discrete Fourier transform and $E(k)$ is discrete Fourier transform (DFT) of $e(n)$ and N is the number of points for computing DFT.

Speech segments of 20 ms and corresponding HE for vowel, semi-vowel and nasal are shown in Fig. 4 (a) - (c) and (d) - (f), respectively. It can be observed that, the pattern of side-lobes of each peak in HE (corresponding to GCI) is different for nasals, semi-vowels and vowels. The side-lobes have higher values with respect to peak values in case of nasals than semi-vowels. In case of vowels, the amplitude of side-lobes are further reduced than that of semi-vowels.

For the entire TIMIT test database, HE of LP residual of vowels, semi-vowels and nasals are obtained. The GCIs are derived from the ZFF signal and then by searching for the nearest peaks in the HE of LP residual [15], [27], [34]. For each GCI, 1.5 ms segment towards right and 1.5 ms segment towards left is selected from the HE of LP residual of speech signal. These 3 ms segments are normalized (each sample is divided by maximum value among the 3 ms samples) and superimposed for each class (vowels, semi-vowels and nasals). The number of such superimposed frames used is equal for each class. The resulting plot is shown in Fig. 5. It can be clearly observed that the distribution of side-lobes around the center peak is different for the three classes of speech sounds.

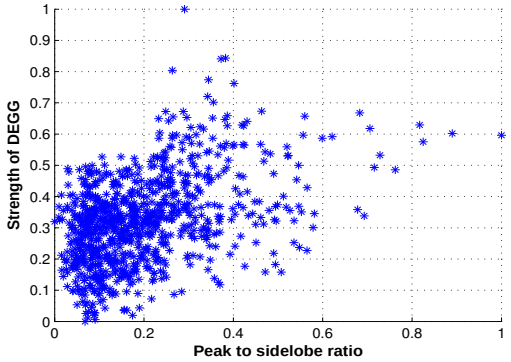

Fig. 7. *Scatter plot of DEGG versus peak to side-lobe ratio of short segment of HE pf LP residual in the vicinity of GCIs*

To investigate the difference among the three, the 3 ms segment is divided further into frames of 0.25 ms. The distribution of values for each 0.25 ms frame is plotted using a discrete histogram as shown in Fig. 6, where, (a), (b), (c), (d) correspond to first 0 to 1 ms (4 frames each of 0.25 ms) and (e), (f), (g), (h) correspond to 2 to 3 ms of 3 ms of HE segment. It can be observed from Fig. 6 that (e), (f), (g), (h) show more discrimination between the classes (vowels, semi-vowels and nasals) than first 1 ms frames i.e. (a), (b), (c), (d). For example: the bins corresponding to vowels, semivowels and nasals are more separated in (f) compared to that in (b). Based on this analysis, we considered only the region from 2 to 3 ms of the 3 ms HE segment to quantify the source evidence. Since the distribution of values of HE of LP residual in glottal closure region is different for broad classes of sonorant sounds (vowels, semi-vowels and nasals), it may be appropriate to analyze the same to quantify the sonority hierarchy.

The source feature for sonority is defined as $f_6 = \frac{P}{\mu}$, where, P is the value of central peak at the GCI location and μ is the mean of sample values from 2 to 3 ms duration in the 3 ms

HE segment. This can be referred as *peak to side-lobe ratio* around the epoch locations which can represent SoE. As shown in Fig. 7, the SoE derived from HE of LP residual (peak to side-lobe ratio) has approximately linear correspondence with strength of DEGG signal. The distribution of peak to side-lobe ratio representing SoE for different classes of sound shows an increasing trend with the increase in sonority which can be observed from Fig. 3(f). The feature of excitation source shows a significant overlap within the vowel categories, whereas it has potential to correctly discriminate source aspect of nasals and vowels. Semi-vowels (glides and liquids) also seem to have overlapped distributions. However, the distributions of f_6 for each class shows less variance compared to that of features of vocal-tract system.

IV. Suprasegmental Evidence for Sonority Measurement

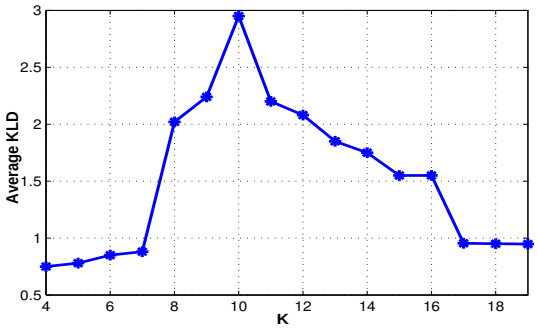

Fig. 8. *Change in average KLD between Gaussian distributions derived from suprasegmental feature of six classes of sonorant sound with respect to the value of K.*

Sonorant sounds are prolonged with higher periodicity, where similar signal structure repeats for longer duration due to the slow change in vocal-tract configuration during production. This behavior of sonorant sounds can be captured by measuring similarity of speech signal samples over several pitch periods rather than just one pitch period. In this work, a suprasegmental feature is derived by computing correlation of speech signal over K pitch periods as a manifestation of regularity in the structure of speech signal. If there are M number of epochs in the given speech signal, $x_1, x_2, ...x_{M-1}$ are the segments corresponding to $M-1$ number of cycles starting from one epoch to the next. The similarity over K number of cycles (pitch periods) is measured as follows:

$$f_7(i) = \frac{1}{K}\sum_{j=i+1}^{i+K} \frac{< x_i, x_j >}{\sum^{N_i} x_i^2 \sum^{N_j} x_j^2}; i = 1, 2, ...M-1-K \tag{11}$$

where, $f_7(i)$ is the correlation coefficient representing suprasegmental evidence of sonorants. $< x_i, x_j >$ represents the inner product between samples corresponding to x_i and x_j, which are i^{th} and j^{th} pitch cycles in the speech segment. Zero padding is performed to match the dimension of x_i and x_j. N_i and N_j are the number of samples present in i^{th} and j^{th} cycles. M is the total number of GCIs in the given speech segment and K is the number of cycles over which the similarity measure is calculated.

For finding appropriate value of K, the suprasegmental feature is derived by varying K value from 4 to 19. For each value of K, Gaussian distributions of the six classes are obtained and average KLD measure among the six classes is calculated. The K value which gives maximum KLD distance between the distribution of six sonorant classes is selected. Figure 8 shows that for $K = 10$, the KLD distance has highest value. If the length of the speech segment is less than 10 pitch periods, the K value is changed to two less than the number of pitch periods in the signal. For M number of GCIs in the speech signal, suprasegmental feature f_7 will have $M-1-K$ number of values. This corresponds to first $M-1-K$ number of epochs. For last $K+1$ number of epochs, the last value of feature is repeated to match the suprasegmental feature dimension with that of vocal-tract system and excitation source feature. The derived correlation feature is obtained for different categories of sonorants from TIMIT test database and the corresponding distribution is depicted in Fig. 3(g). As hypothesized, proposed suprasegmental aspect of speech signal has the adequacy to delineate the sonority hierarchy. Regardless of the significant overlap between distributions of liquids, glides and high-vowels in Fig. 3(g), it shows an increase in feature value as one moves from nasals (least sonorous) to low-vowels (most sonorous).

V. Combination of Source, System and Suprasegmental Evidence

The means and standard deviations of each of the derived features are shown in Table II. As elaborated in Section II-D, the means and standard deviations of five different features of vocal-tract system carry different information regarding the degree of sonority associated with. As observed from Table II, from low-vowels to nasals, the mean values of f_1, f_2, f_3 and f_4 decrease sequentially with a disparity in case of glides and liquids. The latter having higher mean value than the former in case of all the four features. It can be observed that the mean values of f_5 increase from low-vowels to nasals. The deviation in mean values of f_5 among different classes is less. Also, the standard deviation values of f_5 are low compared to other features of vocal tract system.

From production point of view, the difference between glides and liquids is that, in case of liquids the constriction is shorter than that of the glides. This results in higher F_1 for liquids than glides. Moreover, the acoustic path in the oral cavity for liquids contains side branch or parallel paths unlike glides. This introduces extra poles and zeros in the spectrum of liquids which lead to higher values of features of vocal-tract system for liquids than glides. The pattern of mean values of suprasegmental feature is found to have good correlation with the degree of sonority. All the evidences derived from three different perspectives of sonorant sounds demonstrate unique trend with the change in degree of sonority. To obtain a faithful feature representation of sonority, the combination of features of vocal-tract system, feature of excitation source and suprasegmental feature may be helpful. All the seven evidences have one value at each epoch location.

For each of the seven features, six Gaussian distributions can be derived representing six classes of sonorant sounds. The

TABLE II
Means and standard deviations (std) of different features of vocal-tract system (f_1, f_2, f_3, f_4, f_5), feature of excitation source (f_6) and suprasegmental feature (f_7) for different classes of sonorants (low-vowel, mid-vowel, high-vowel, liquid, glide and nasal).

Evidence	Low-vowel		Mid-vowel		High-vowel		Glide		Liquid		Nasal	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Formant Peak Values (f_1)	0.73	0.11	0.69	0.12	0.56	0.12	0.48	0.13	0.62	0.14	0.32	0.09
Formant peak deviation (f_2)	0.60	0.14	0.56	0.14	0.54	0.12	0.46	0.11	0.53	0.14	0.38	0.08
Spectral valleys (f_3)	0.62	0.12	0.59	0.12	0.49	0.13	0.45	0.13	0.55	0.14	0.33	0.09
Slope (f_4)	0.71	0.12	0.67	0.12	0.54	0.11	0.46	0.12	0.60	0.14	0.29	0.09
Formant Bandwidth (f_5)	0.55	0.05	0.58	0.05	0.57	0.05	0.59	0.05	0.61	0.06	0.63	0.06
Source(f_6)	0.29	0.06	0.29	0.06	0.29	0.06	0.24	0.08	0.27	0.08	0.20	0.08
Suprasegmental(f_7)	0.49	0.14	0.44	0.15	0.34	0.16	0.32	0.15	0.29	0.14	0.24	0.11

distance between each pair of Gaussian probability density function can be measured by Kullback Leibler divergence (KLD) [35] as given by (12).

$$D_{KL}(A,B) = \frac{1}{2}\left\{\frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2}\right\} - 1 + \frac{1}{2}\{\mu_A - \mu_B\}^2\left\{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}\right\} \quad (12)$$

where, A and B are two univariate Gaussian distributions with mean μ_A, μ_B and standard deviation σ_A, σ_B, respectively. Here A and B represent samples of one feature for two classes of sonorant sounds. As there are 6 classes of sonorant sounds, each feature will have 6 Gaussian distributions i.e. 15 pairs of distributions as shown in Fig. 3. The average KLD distance measure is calculated for each of the seven features over these 15 pairs of distribution as in (13). The average KLD distance for each feature is tabulated in Table III.

$$\{D_{KL}(A,B)\}_{avg} = \frac{1}{15}\sum_{i=1}^{15} D_{KL}(A,B)_i \quad (13)$$

The seven features shown in Table III have difference in terms of their ability to differentiate between the classes of sonorant sounds. High value of KLD represents greater ability of the feature to discriminate different classes of sonorants and hence more weight should be assigned to that particular feature dimension. Based on the average KLD between different classes of sound, weights corresponding to each of the seven features (w_i) are derived such that

$$\sum_{i=1}^{7} w_i = \quad 1 \quad (14)$$

where,

$$w_i = \quad \frac{[\{D_{KL}(A,B)\}_{avg}]_{f_i}}{\sum_{i=1}^{7}[\{D_{KL}(A,B)\}_{avg}]_{f_i}} \quad (15)$$

The weights assigned to each of the seven features according to their potential to classify different sonorant sounds are also shown in Table III. Thus a competent representation of degree of sonority associated with a sound unit is derived in this work.

The overall block diagram of the proposed work is depicted in Fig. 9. Three different features are derived using the knowledge of vocal-tract system, excitation source and suprasegmental aspects of sonorants. To derive the feature

TABLE III
Average KLD between Gaussian distributions of six classes of sonorant sounds and corresponding weights assigned for different features of vocal-tract system, exciation source and suprasegmental feature.

Features	Average KLD	Weights
Formant Peak Values (f_1)	1.14	0.1049
Formant peak deviation (f_2)	0.95	0.0874
Spectral valleys (f_3)	1.10	0.1012
Slope (f_4)	1.09	0.1003
Formant Bandwidth (f_5)	1.62	0.1490
Source (f_6)	2.02	0.1858
Suprasegmental (f_7)	2.95	0.2714

of vocal-tract system, ZTW is performed around each epoch location of speech signal. For the windowed segments, HNGD spectra are derived. Feature of excitation source is derived from the HE of LP residual of speech signal. In contrast to these two evidences, the suprasegmental feature is derived from correlation of speech signal over ten pitch periods. The three evidences are weighted and fused together to derive the seven-dimensional sonority evidence (vocal-tract system (five-dimension), excitation source (one-dimension) and suprasegmental feature (one-dimension)). The implementation for extraction of this sonority feature is released in the following link [1]. The evidence is further utilized in the task of sonorant/non-sonorant classification, multiclass sonorant classification and phoneme recognition to verify the efficacy of the proposed feature.

VI. Experimental Evaluation

TABLE IV
Comparison of performance of proposed feature (using SVM) and existing feature using hierarchical algorithm (within braces) as shown in [23] in sonorant/non-sonorant segmentation on utterances from TIMIT database in both clean speech and noisy speech across different SNR levels.

SNR	Proposed feature (Existing Feature)					
	Epoch based results			Frame based results		
	Acc(%)	TPR(%)	FAR(%)	Acc(%)	TPR(%)	FAR(%)
clean	96.3 (93.9)	98.5 (94.5)	5.5 (7.5)	95.0 (92.8)	97.3 (93.6)	6.8 (8.0)
30 dB	96.0 (93.9)	98.4 (94.5)	5.7 (7.6)	94.8 (92.8)	96.7 (93.6)	7.6 (8.0)
20 dB	95.4 (94.39)	96.6 (94.4)	6.2 (7.7)	93.5 (92.7)	95.8 (93.5)	8.0 (8.1)
10 dB	95.0 (93.4)	94.3 (94.0)	6.8 (8.5)	93.4 (92.1)	95.3 (92.9)	8.4 (9.0)
5 dB	93.4 (92.4)	93.6 (93.1)	8.3 (8.9)	93.0 (91.0)	92.8 (91.9)	9.3 (9.5)
0 dB	90.5 (90.7)	91.4 (91.0)	9.5 (9.9)	90.0 (89.6)	90.7 (89.9)	10.3 (10.6)

[1] https://github.com/bidishasharma/Extract-Sonority-Feature

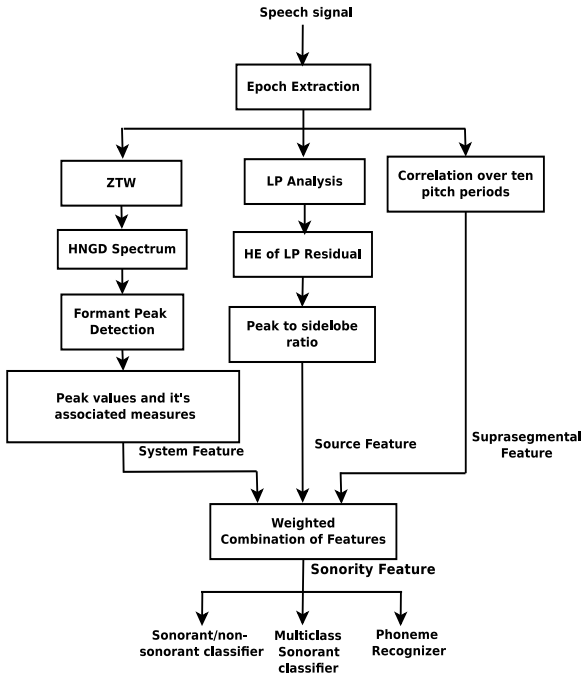

Fig. 9. *Overall block diagram of the proposed sonority feature extraction from speech signal, where vocal-tract system, excitation source and suprasegmental features are derived from HNGD spectrum, HE of LP residual and speech signal, respectively. These features are combined to derive the sonority feature.*

The distribution of the proposed sonority evidence correlates well with the sonority hierarchy as can be observed from Fig. 3 and Table II. To establish the efficacy of the proposed seven-dimensional sonority feature vector in different speech processing applications, the following classification experiments are performed.

A. Sonorant/non-sonorant classification

The first level of classification that exploits the usefulness of prospective features representing sonority is sonorant/non-sonorant classification. In [23], it has been demonstrated that the attributes derived from speech signal like zero frequency resonator (ZFR) signal energy, slope of ZFR signal around epoch locations and dominant resonance frequency (DRF), can be used for the task of sonorant/non-sonorant segmentation, both at frame and epoch levels. An hierarchical algorithm is used for the classification task. To compare the effectiveness of the proposed feature with the features used in [23], a sonorant/non-sonorant classifier using support-vector-machine (SVM) (with radial basis function (RBF) kernel, $c = 16$, $\gamma = 4$) is developed using the proposed sonority feature vector. The training and testing feature vectors are derived from all SI and SX utterances of TIMIT train and test databases, respectively. This is followed by feature normalization to make the feature values within zero to one range. Similar normalization is performed in training and testing of clean and noisy speech. The same SVM classifier trained using clean speech is employed in the testing of sentences mixed with white noise across various SNR levels.

To demonstrate the robustness of the features for classification, the performance evaluation parameters used are: number of epochs/frames correctly detected in the sonorant regions (true positive rate (TPR)), number of spurious epochs/frames hypothesized in the non-sonorant regions (false alarm rate (FAR)) and total number of correctly detected epochs/frames in both the sonorant and non-sonorant regions (accuracy (Acc)). As shown in Table IV, the proposed feature can segment sonorant regions with more accuracy compared to the existing method (within braces). Table IV shows that the proposed feature has better ability to classify sonorant/non-sonorant segments from the given speech signal.

B. Classification of sonorant sounds into different classes

The primary motivation of this work is to derive feature to characterize the degree of sonority associated with a sound unit. The straightforward way to validate this would be to develop a multi-class sonorant classifier. Each class represents different sonorant sounds (low-vowels, mid-vowels, high-vowels, liquids, glides and nasals). As described in Section V, the proposed seven-dimensional sonority feature is derived for each class of sonorant sounds for the entire TIMIT test database. This is followed by normalization to make the feature value within the range of 0 to 1. Individual feature dimension consists of a single value at each epoch location. A six-class SVM classifier (with RBF kernel, $c = 256$, $\gamma = 16$) has been developed using the normalized sonority feature vector. Values of parameters, c and γ are set using train-test 5-fold cross validation for the entire TIMIT test database. For the optimized value of c and γ, the six-class SVM model is trained using randomly chosen 80% of TIMIT-test data. The rest 20% data is used for testing.

The classification accuracy of each class and confusion among different classes are reported in Table V. The average accuracy achieved is 66.55%. The accuracy is observed to be the lowest for liquids and highest for nasals. It can be interpreted from Table V that, 14.41% of low-vowels are misclassified as mid-vowels. This is due to the fact that the properties of low-vowels and mid-vowels are close to each other. Moreover, as observed from Fig. 3, formant bandwidth and feature of excitation source exhibit overlap between the two classes. As the height of the tongue body for mid-vowels is intermediate between that of high and low-vowels, it affects the constriction size and length. This in-turn alters the VTS evidences.

Although the vocal-tract constriction in case of liquids is narrower than the glides resulting in wider F_1 bandwidth for liquids, the length of constriction is shorter in case of liquids. This increases F_1 for liquids and introduces confusion between glides and liquids. Thus there is possibility of confusion of liquids with low-vowels and mid-vowels. This is evident from 1^{st}, 2^{nd} and 5^{th} rows of Table V. The common attribute of liquids with vowels is that, in both cases air flows through the constriction without pressure drop. As a result, the vocal-folds continue to vibrate in the period of constriction. In the distribution of feature of excitation source in Fig. 3(f), confusion between glides and liquids can be apparently observed. As reported in Table V, majority of misclassification of high-vowels is due to the confusion with mid-vowels and glides.

The configuration of vocal-tract for glides may also change based on the preceding vowels. A glide adjacent to high-vowel is produced with more constricted structure compared to the one preceded or followed by a low-vowel. Therefore, when a glide is contiguous with low-vowel or mid-vowel, due to less constriction, F_1 may increase. The bandwidth may decrease compared to the glide that is adjacent with a high-vowel.

The proposed features are analogous to formant based measures and do not use the temporal information of nearby sounds. Therefore, there is a possibility of misclassification of each category to its adjacent category of sound in the sonority hierarchy. It is notable from Fig. 3 that, compared to other categories of sonorants, the distribution corresponding to nasals has less overlap with other distributions. Only in case of suprasegmental feature in Fig. 3(g), some confusion with nasals and other categories is observable. This correlates with highest accuracy for nasals as reported in Table V. As the front part of vocal-tract is completely closed during nasal murmur, the first formant frequency and its prominence eventually decreases with a weak second formant followed by an extended valley in the VTS. This is more contrasting with other sonorants. However, the common acoustic behavior of nasals and glides is that, the vocal-fold does not change the vibration pattern before and after the constriction happens. Based on this discussion and the classification accuracy of sonorants presented in Table V, it can be inferred that the proposed features have ability to quantify sonority level associated with a sound unit. Although, some aspects of the speech signal corresponding to a specific category of sound unit may vary based on the adjacent sound units present.

TABLE V
Classification accuracy (epoch level) of different sonorant sounds from TIMIT test database using SVM ($c = 256, \gamma = 16$) obtained by employing the proposed seven-dimensional sonority feature

Category	% Accuracy					
	Low-vowel	Mid-vowel	High-vowel	Glide	Liquid	Nasal
Low-vowel	**68.0**	14.4	4.1	2.8	9.2	1.5
Mid-vowel	9.8	**63.9**	9.2	4.5	10.9	1.7
High-vowel	1.7	10.3	**67.3**	11.7	4.6	4.4
Glide	1.4	6.4	12.7	**59.4**	6.7	13.4
Liquid	7.2	13.3	9.9	8.5	**55.9**	5.2
Nasal	0.5	1.9	3.4	1.5	7.9	**84.8**

To further demonstrate the ability of the proposed features for discriminating different sonorant classes, in addition to MFCC, two SVM classifiers (one using sonority feature and the other using MFCC feature) are fused at score level [36]. For this thirteen-dimensional MFCC feature is used to develop another six class SVM classifier (with RBF kernel, $c = 2, \gamma = 4$), where c and γ values are set using train-test 5-fold cross validation for entire TIMIT test database. For the optimized values of c and γ, the six-class SVM model is trained. The randomly chosen 80% of TIMIT-test data is used for training and rest 20% is used for testing. The average accuracy of the MFCC based classifier is found to be 80.41%. The detailed performance for each class can be seen in Table VI (within braces). As there are 6 classes, each of the classifiers using MFCC and sonority feature will produce 6 posterior probabilities for each feature vector.

For the sonority based classifier, the posterior probability scores corresponding to epochs within one frame are averaged to derive single probability score corresponding to each class for each frame. The mean value of probabilities of the two classifiers for each class corresponding to each frame is calculated to derive the fused probability score. The class with maximum average probability score is considered as final output of the combined classifier. The resultant accuracy of the combined classifier is found to be 84.51%, which is 80.41% when only MFCC feature is used. The classification accuracy for each class using the combined classifier and only MFCC based classifier is shown in Table VI for comparison. By comparing both % accuracy values in Table VI, an absolute improvement of 4.1% can be observed when the two classifiers are fused. For each of the classes, along with improvement in classification, reduction in confusion among different sonorant classes can also be observed. It is interesting to observe from Table VI that, with increase in correct classification of each class, the percentage of confusion with other classes is reduced for most of the cases.

To study individual performances of sonorant classification for male and female, we have developed two sonorant classifiers using SVM (with RBF kernel, $c = 256$, $\gamma = 16$) for male and female utterances from TIMIT test database. For developing each classifier 80% of male/female data is used for training and rest 20% is used for testing. The average accuracy of the six class sonorant classification is found to be is 68.4% for male and 65.6% for female. The relatively poor performance for the female case may be attributed to the associated high non-stationarity nature.

C. Effect of noise on sonority feature

In order to analyze the impact of noise on the proposed features, the classifier trained using features derived from clean speech is employed for testing of noisy cases. The test features are derived after addition of different kinds of noises (babble noise, factory noise, white noise) to the speech signal at different SNR levels (0 dB, 5 dB, 10 dB, 15 dB). The average accuracy the six classes for different types and levels of noise is shown as bar plot in Fig. 10. It can be observed that % accuracy significantly decreases in case of 0 dB and 5 dB SNR levels. Whereas, for 10 dB and 15 dB cases, % accuracy is less effected. Further, to analyze the robustness of each of the system, source and suprasegmental features, three six-class SVM classifiers are developed using individual features derived from clean speech. The test features are derived after adding different levels of babble noise with the speech signal.

Figure 11 demonstrates degradation of % accuracy of the three classifiers with increased noise level. This depicts that the suprasegmental feature is more affected due to noise compared to the features of vocal-tract system and excitation source. This may be due to the reason that, suprasegmental feature is directly derived from the speech signal by measuring correlation over successive pitch periods. Furthermore, it is not derived in synchrony with glottal closed phase which may be less susceptible to degradation due to noise. The features of vocal-tract system are derived from HNGD spectrum which is

TABLE VI

Classification accuracy of different sonorant segments (frame level) from TIMIT database using combined sonority and MFCC feature based SVM classifier. Classification accuracy obtained by using only MFCC feature vector is shown within braces ($c = 2, \gamma = 4$)

Category	% Accuracy					
	Low-vowel	**Mid-vowel**	**High-vowel**	**Glide**	**Liquid**	**Nasal**
Low-vowel	**86.3 (78.7)**	6.5 (13.6)	3.2 (3.4)	0.2 (0.4)	3.8 (2.7)	0.0 (0.9)
Mid-vowel	10.8 (10.4)	**75.4 (68.8)**	5.3 (11.4)	0.7 (1.6)	7.8 (7.3)	0.0 (0.5)
High-vowel	0.5 (0.4)	7.3 (9.6)	**85.2 (80.8)**	5.8 (5.1)	0.9 (3.1)	0.3 (0.4)
Glide	0.1 (0.3)	2.0 (1.1)	6.8 (8.8)	**83.5 (80.7)**	5.4 (5.3)	2.2 (3.8)
Liquid	3.6 (4.1)	6.5 (5.3)	1.5 (2.8)	5.8 (4.5)	**80.7 (78.8)**	1.9 (4.5)
Nasal	0.2 (0.2)	0.8 (0.5)	0.8 (1.8)	0.9 (1.2)	1.3 (1.7)	**96.0 (94.7)**

reported to be less affected by different types of noise [14]. This happens due to the short and tapered window used in HNGD. For deriving feature of excitation source, the samples corresponding to glottal closed phase around epoch locations is accessed. Hence this feature is also found to be less affected by noise.

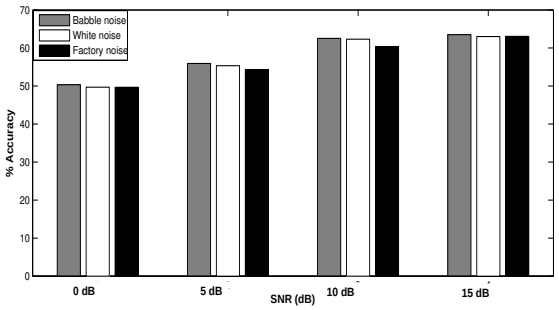

Fig. 10. *Bar plot representing average % accuracy for SVM based six-class sonorant segment classification in presence of different types of noise with different SNR levels.*

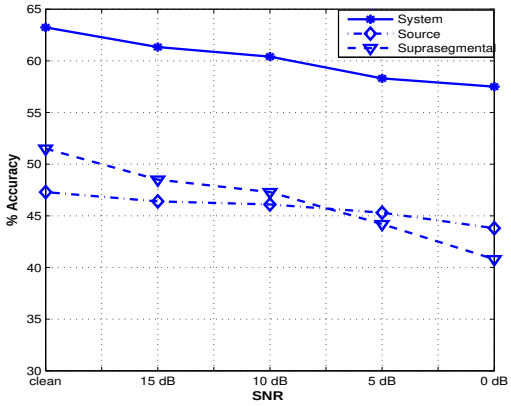

Fig. 11. *Average % accuracy of six-class sonorant classifier using each of the system, source and suprasegmental features in with respect to different levels of noise.*

The above experiments validate the effectiveness of the proposed feature in discriminating the sonorant sounds or characterization of degree of sonority from given speech signal. To show its usefulness in different speech processing applications, the proposed sonority feature is used in phoneme recognition.

D. Sonority as a feature for phoneme recognizer

The proposed sonority feature may also be helpful to improve the performance of a phoneme recognizer by incorporating additional information to reduce confusion among different sonorants. In this regard, phoneme recognition framework for TIMIT database is developed in Kaldi toolkit [37], [38], where deep neural network (DNN) based acoustic modeling is implemented [39]. In addition to traditional MFCC feature, proposed seven-dimensional weighted sonority feature is employed for developing the recognizer. The proposed feature is epoch synchronous. In order to match dimension with MFCC at frame level, average value of feature corresponding to epochs within one frame is calculated. It is then appended with the thirteen-dimensional MFCC feature resulting in a twenty-dimension feature vector. A bigram phoneme language model created from the training set is incorporated in the recognizer.

The 61 phonemes are mapped into 39 phonemes for training and testing, the acoustic model is an HMM-DNN hybrid model. The training set contains $3,696$ sentences from 462 speakers. The development set contains 400 sentences from 50 speakers. Core test set is also used as test set, which contains 192 sentences from 24 speakers. The number of hidden layers used is 2. It is reported in Kaldi documentation that 4 hidden layers are effective when 100 hours of speech data is available. An initial learning rate of 0.015 is selected which is reduced to 0.002 in 20 epochs. Additional 10 epochs are employed after reducing the learning rate to 0.002. Kaldi employs a preconditioned form of stochastic gradient descent. A matrix-valued learning rate is employed instead of using a scalar learning rate in order to reduce the learning rate in dimensions where the derivatives have a high variance. This is in order to control instability and stop the parameters moving too fast in any one direction.

The overall performance of the baseline phoneme recognizer using MFCC as feature and using additional proposed feature (MFCC + sonority) is shown in Table VII in terms phone error rate (% PER). It is improved while using proposed features along with MFCC. Also, the improvement in case of different sonorant phones in terms of accuracy (%) and correct (%) identification is shown in the bar plot of Fig. 12. The performance increases after using the proposed sonority features. It is observed that with the addition of proposed evidence, insertion and substitution of sonorant phones decreases significantly, whereas the reduction in deletion is comparatively less. However, the confusion among different classes of sonorant phones is analyzed in terms of % substitution. It seems to

reduce while employing the proposed feature in addition to MFCC as shown in Table VIII.

TABLE VII
Phone error rate (PER) for DNN based phoneme recognizer by using MFCC and (MFCC+Sonority) feature

Evaluation on	PER(%)	
	MFCC	MFCC+sonority feature
Test set	22.7	21.4
Dev set	21.2	20.3

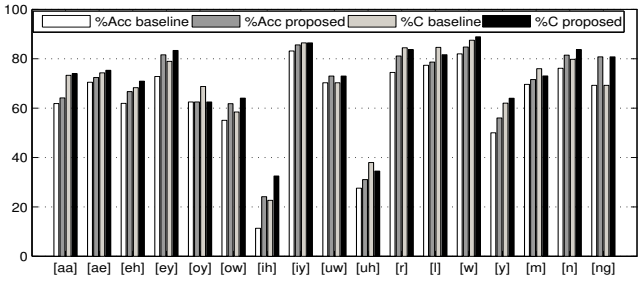

Fig. 12. *Correction percentage (%C) and accuracy (%Acc), before and after appending the sonority for various sonorant phones of TIMIT.*

TABLE VIII
% substitution of different sonorant phones before and after appending the proposed sonority evidence for various sonorant phones of TIMIT. Baseline result using MFCC is shown braces.

Category	% Substitution						
	Low-vowel	Mid-vowel	High-vowel	Glide	Liquid	Nasal	Total
Low-vowel	4.0 (4.1)	5.5 (6.1)	6.8 (7.0)	0.0 (0.1)	0.5 (0.5)	0.3 (0.4)	**17.1 (18.2)**
Mid-vowel	3.5 (4.9)	1.3 (1.3)	4.8 (4.9)	0.0 (0.0)	0.8 (1.2)	0.1 (0.1)	**10.5 (12.4)**
High-vowel	4.3 (4.7)	1.9 (2.1)	4.7 (5.3)	0.4 (0.9)	0.2 (0.2)	0.2 (0.7)	**11.7 (13.9)**
Glide	0.0 (0.0)	0.0 (0.0)	1.0 (1.8)	0.3 (0.3)	0.3 (0.5)	0.5 (0.8)	**2.1 (3.4)**
Liquid	0.6 (1.2)	0.8 (0.9)	0.3 (0.2)	0.1 (0.3)	0.1 (0.1)	0.4 (0.7)	**2.3 (3.4)**
Nasal	0.5 (0.5)	0.2 (0.3)	0.5 (0.5)	0.1 (0.2)	0.2 (0.3)	3.6 (3.9)	**5.1 (5.7)**

VII. Summary, Conclusions and Scope

In this work, an effort is made to define a feature which can represent the degree of sonority associated with a sound unit. For this task, different characteristics of sonorant sounds reflected in the speech signal are analyzed. Consequently features based on vocal-tract system, excitation source and suprasegmental aspects are derived. These features correlate with less vocal-tract constriction, glottal vibration and periodicity properties of sonorant sounds. To justify, whether each of the proposed features can represent the level of sonority, distributions for feature values are shown for different sonorant sounds along the sonority hierarchy. Each of the proposed features shows increasing/decreasing trend in feature value with the increase in sonority. The proposed seven-dimensional sonority feature is used in classification among different sonorant sounds and is found to be potential for the same. It is also shown to be useful for the phoneme recognition application. In future we may focus on exploring evidences which can reduce the confusion among adjacent classes in the sonority hierarchy.

VIII. Acknowledgement

This work is a part of the ongoing project on the "Development of Text-to-Speech Synthesis for Assamese and Manipuri languages" funded by TDIL, DEiTy, MCIT, GOI. The authors would also like to thank Mr. Abhishek Dey for his kind help in developing DNN based phoneme recognition framework.

References

[1] S. G. Parker, "Quantifying the sonority hierarchy," Ph.D. dissertation, University of Massachusetts Amherst [Published by the GLSA.], 2002. 1

[2] K. Schutte and J. R. Glass, "Robust detection of sonorant landmarks." in *INTERSPEECH*, 2005, pp. 1005–1008. 1, 2

[3] S. Parker, "Sound level protrusions as physical correlates of sonority," *Journal of phonetics*, vol. 36, no. 1, pp. 55–90, 2008. 1

[4] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999. 1

[5] J. Blevins and J. Goldsmith, "The syllable in phonological theory," *Phonology: Critical Concepts: Syllables and Multi-level Analyses*, vol. 3, pp. 75–120, 2001. 1

[6] M. Gouskova, "Relational hierarchies in optimality theory: the case of syllable contact," *Phonology*, vol. 21, no. 02, pp. 201–250, 2004. 1

[7] P. De Lacy, "Markedness conflation in optimality theory," *Phonology*, vol. 21, no. 02, pp. 145–199, 2004. 1

[8] E. Moreton, G. Feng, and J. L. Smith, "Syllabification, sonority, and perception: new evidence from a language game," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, vol. 41, no. 1. Chic Ling Society, 2005, pp. 341–355. 1

[9] S. Topbas and H. KopkallI-Yavuz, "Reviewing sonority for word-final sonorant+ obstruent consonant cluster development in turkish," *Clinical linguistics & phonetics*, vol. 22, no. 10-11, pp. 871–880, 2008. 1

[10] M. E. Beckman, J. Edwards, and J. Fletcher, "Prosodic structure and tempo in a sonority model of articulatory dynamics," *Papers in laboratory phonology II*, pp. 68–86, 1992. 1

[11] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Communication*, vol. 1, no. 3, pp. 167–184, 1982. 1

[12] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering,*, vol. 41, no. 7, pp. 663–671, 1994. 1

[13] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Transactions on Speech and Audio Processing,*, vol. 6, no. 4, pp. 313–327, 1998. 2

[14] B. Yegnanarayana and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013. 2, 3, 4, 5, 12

[15] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 16, no. 8, pp. 1602–1613, 2008. 2, 3, 4, 7

[16] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012. 2

[17] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2061–2071, 2009. 2

[18] S. Puppel, "The sonority hierarchy in a source-filter dependency framework," in *Phonological investigations (Linguistic & Literary Studies in Eastern Europe, volume 38.)*, J. Fisiak and S. Puppel, Eds. Amsterdam and Philadelphia: John Benjamins Publishing Company, 1992, pp. 467–483. 2

[19] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *ICSLP*, vol. 2. IEEE, 1996, pp. 1261–1264. 2

[20] J. Yuan and M. Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *ICASSP*. IEEE, 2010, pp. 4222–4225. 2

[21] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 15, no. 8, pp. 2190–2201, 2007. 2

[22] A. R. Arrabothu, N. Chennupati, and B. Yegnanarayana, "Syllable nuclei detection using perceptually significant features." in *INTERSPEECH*, 2013, pp. 963–967. 2

[23] S. H. Dumpala, B. T. Nellore, R. R. Nevali, S. V. Gangashetty, and B. Yegnanarayana, "Robust features for sonorant segmentation in continuous speech," in *INTERSPEECH*, 2015. 2, 9, 10

[24] A. Jansen and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1739–1758, 2008. 2

[25] B. D. Sarma, B. Sharma, S. A. Shanmugam, S. M. Prasanna, and H. A. Murthy, "Exploration of vowel onset and offset points for hybrid speech segmentation," in *TENCON 2015-2015 IEEE Region 10 Conference*. IEEE, 2015, pp. 1–6. 2

[26] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30. 3

[27] B. Sharma and S. M. Prasanna, "Speech synthesis in noisy environment by enhancing strength of excitation and formant prominence." in *INTERSPEECH*, 2016, pp. 131–135. 3, 7

[28] B. Sharma and S. Prasanna, "Faster prosody modification using time scaling of epochs," in *2014 Annual IEEE India Conference (INDICON)*. IEEE, 2014, pp. 1–5. 3

[29] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1638–1640, 1978. 4

[30] G. A. Seber, *Multivariate observations*. John Wiley & Sons, 2009, vol. 252. 6

[31] J. R. Schott, "Principles of multivariate analysis: A user's perspective," *Journal of the American Statistical Association*, 2011. 6

[32] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975. 6

[33] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech and Signal Processing,*, vol. 27, no. 4, pp. 309–319, 1979. 7

[34] B. Sharma and S. M. Prasanna, "Improvement of syllable based tts system in Assamese using prosody modification," in *2015 Annual IEEE India Conference (INDICON)*. IEEE, 2015, pp. 1–6. 7

[35] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012. 9

[36] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998. 11

[37] Kaldi Toolkit: http://kaldi.sourceforge.net. 12

[38] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. ASRU*, Dec 2011. 12

[39] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012. 12

Bidisha Sharma received B.E. degree in Electronics and Telecommunication Engineering from Girijananda Chowdhury Institute of Management and Technology (GIMT), Gauhati University, Guwahati, India, in 2012. She is currently pursuing PhD in the Department of Electronics and Electrical Engineering, Indian Institute of Technology (IIT) Guwahati. Her research interests are in speech signal processing, speech synthesis and speech enhancement.

S. R. Mahadeva Prasanna received the B.E. degree in Electronics Engineering from Sri Siddartha Institute of Technology, Bangalore University, Bangalore, India, in 1994, the M.Tech. degree in Industrial Electronics from the National Institute of Technology Karnataka (NITK), Surathkal, India, in 1997, and the Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology (IIT) Madras, Chennai, India, in 2004. He is currently a Professor in the Department of Electronics and Electrical Engineering, Indian Institute of Technology (IIT) Guwahati. His research interests are in speech, handwriting and signal processing.